D824–D831 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 5 November 2008*
doi:10.1093/nar/gkn832

MoKCa database—mutations of kinases in cancer

Christopher J. Richardson[1], Qiong Gao[2], Costas Mitsopoulous[2], Marketa Zvelebil[2], Laurence H. Pearl[1] and Frances M. G. Pearl[1,*]

[1]Section of Structural Biology and [2]The Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham Road, London SW3 6JB, UK

Received August 14, 2008; Revised October 3, 2008; Accepted October 14, 2008

ABSTRACT

Members of the protein kinase family are amongst the most commonly mutated genes in human cancer, and both mutated and activated protein kinases have proved to be tractable targets for the development of new anticancer therapies The MoKCa database (Mutations of Kinases in Cancer, http://strubiol.icr.ac.uk/extra/mokca) has been developed to structurally and functionally annotate, and where possible predict, the phenotypic consequences of mutations in protein kinases implicated in cancer. Somatic mutation data from tumours and tumour cell lines have been mapped onto the crystal structures of the affected protein domains. Positions of the mutated amino-acids are highlighted on a sequence-based domain pictogram, as well as a 3D-image of the protein structure, and in a molecular graphics package, integrated for interactive viewing. The data associated with each mutation is presented in the Web interface, along with expert annotation of the detailed molecular functional implications of the mutation. Proteins are linked to functional annotation resources and are annotated with structural and functional features such as domains and phosphorylation sites. MoKCa aims to provide assessments available from multiple sources and algorithms for each potential cancer-associated mutation, and present these together in a consistent and coherent fashion to facilitate authoritative annotation by cancer biologists and structural biologists, directly involved in the generation and analysis of new mutational data.

INTRODUCTION

Cancers arise due to the accumulation of mutations in critical target genes that confer a selective advantage on the cell (and its progeny) that contain them. Knowledge of these mutations is key to understanding the biology of cancer initiation and progression, as well as to the development of more targeted therapeutic strategies. While there are rare examples of cancers driven by a single genetic alteration (1), in most solid tumours, tumourigenesis is a multistep process (2,3), reflecting the genetic alterations necessary to transform normal cells into malignant derivatives. The crucial events include acquisition of genomic instability, cell cycle deregulation, evasion of apoptosis, limitless replicative potential, angiogenesis and metastasis (4). Regardless of the multiple mutations ultimately required, a single mutation can initiate the process.

Members of the protein kinase family are amongst the most commonly mutated genes in human cancer, and both mutated and activated protein kinases have proved to be tractable targets for the development of new anticancer therapies (5). There are 518 documented mammalian protein kinases (6) encoded in the human genome, which together represent the largest family of human enzymes, collectively termed the kinome. They play indispensable roles in numerous cellular, metabolic and signalling pathways, in all cell types. Around 40% of the kinome have multiple splice variants and 10% of the total encode catalytically deficient enzymes that have been termed pseudokinases.

Although there are several kinase classification schemes (e.g. 6,7), the KinBase resource (http://www.kinase.com/kinbase) (6) reflects the currently accepted classification of eukaryotic protein kinases where the kinases are broadly split into two groups: conventional protein kinases (ePKs) and atypical protein kinases (aPKs). The ePKs are the largest group, and are subclassified into eight families using the sequence similarity of the kinase domain, the presence of accessory domains, and consideration of their modes of regulation. The aPKs are a smaller set of protein kinases that do not share clear sequence similarity with ePKs, but have been shown experimentally to have protein kinase activity. As the entries in KinBase are filtered by stringent criteria, including verification by

*To whom correspondence should be addressed. Tel: +44 (0)20 7153 5443; Fax: +44 (0)20 7153 5457; Email: frances.pearl@icr.ac.uk

cDNA cloning, the KinBase classification scheme is the one favoured by experimentalists working on kinases and signal transduction pathways.

Protein kinases are frequently found to be mis-regulated in human cancer, and the Cancer Genome Project and similar initiatives, have undertaken systematic re-sequencing screens of all annotated protein kinases in the human genome, to attempt to identify commonly occurring mutations that may play significant roles in a range of different cancers (8–10). In all cases the key to understanding the contribution of a particular disease-associated kinase mutation to development and progression of cancer, comes from an appreciation of the consequences of that mutation on the function of the affected protein, and the impact on the pathways in which that protein is involved. It is this that the MoKCa database described here, aims to facilitate.

Changes in the nucleotide sequence of a gene can have a variety of consequences for the encoded protein, including truncations and frameshifts that disrupt the protein structure and/or reduce transcript levels via nonsense-mediated RNA decay. The most common mutational event is a single base change, leading to a missense mutation of a single amino acid. Cancer-associated somatic mutations (CASMs) or missense variants are commonly identified in somatic tumour DNA, but only a fraction directly contribute to oncogenesis. Distinguishing those that contribute to cancer from those that do not is a difficult problem, potentially requiring detailed and protracted functional analysis. However the ability to make this determination, both rapidly and inexpensively, will be essential to realization of 'personalized therapy' targeted to the individual tumour.

Several approaches have been taken to predict which genes contain mutations that contribute to mutagenesis (i.e. drivers genes) rather than those genes that contain mutations that arise by chance but have no bearing on the disease (passenger mutations). Statistical models comparing the observed ratio of non-synonymous:synonymous compared with that expected by chance, have been used to identify and estimate the number of cancer drivers within a total set of identified genetic variation (11). These methods are excellent for predicting which genes contain drivers, but do not identify the driver mutation alone. Consequently algorithms have been developed that attempt to assess the driver status of missense mutations (12–14) based on the notion that evolutionary conserved sites in a protein tend to be involved in its function, and that mutations that change the properties of these sites, alter that function. However the mechanistic nature of a functional change (activation, inhibition or subversion) and its detailed biochemical effect, is virtually impossible to predict without a detailed analysis of the protein, informed by 'expert' insight into its individual biology.

There are several protein family-specific databases that collate disease-associated mutations from the literature, such as SH2base (15) that contains data for germline mutations in proteins containing SH2 domains, and KinMutBase (16) which documents disease-related germline mutations in 33 kinases. These data either derive from repetitious sequencing in affected families of particular kinases mutated in specific diseases, or from harvesting literature identifications of mutations observed in individual studies. Furthermore, the COSMIC database (17), is undertaking to document all somatic cancer mutations reported in the literature.

By bringing together automatic assessments available from multiple sources and algorithms for each potential cancer-associated mutation, and presenting these in a convenient and coherent fashion, MoKCa aims to facilitate authoritative annotation by cancer biologists and structural biologists, directly involved in the generation and analysis of new mutational data. These 'experts' are then able to bring detailed insights into the biochemistry and biology of individual proteins and systems, that are virtually impossible to encapsulate in an algorithm, but are key to determining if and how a particular mutation will alter the biological function of a protein. Thus the MoKCa database combines automated and 'expert' annotation of individual mutations, and is firmly directed towards the specific needs of the cancer research community.

BUILDING THE MoKCa DATABASE

Mutation data

The original mutational data was provided by the Sanger Cancer Genome Project (CGP) Team, and comprises the mutations found in the large-scale re-sequencing of the kinase complement from 210 human cancers. This included samples from breast, lung, colorectal, gastric, testis, ovarian, renal, melanoma, glioma and acute lymphoblastic leukaemia (ALL) cancers (9). Of the 210 tumours studied, 169 were primary tumours, 2 were early cultures and 39 immortal cell lines.

One-thousand and seven somatic mutations are documented in the coding exons and splice junctions in the kinases of 137 of the tumours studied. Nine-hundred and twenty one were single base substitutions, 78 were small insertions or deletions and 8 were complex changes. Of the single base substitutions, 620 were missense changes, 54 caused nonsense changes and 28 were observed at highly conserved positions of splice junctions. There were also 219 silent (synonymous mutations). Approximately one-third of these mutations had previously been reported in the literature.

Added to this core dataset, are additional somatic mutational data from the COSMIC database that have been curated from the literature (17). This includes 15 911 missense, 47 nonsense and 50 insertions or deletions. This results in a non-redundant mutation dataset of 1406 distinct mutations from over 20 different types of cancer: 269 silent, 912 missense, 83 nonsense, 27 splice site, 84 deletions and 8 multi substitutions.

Driver/passenger assignment of a gene

The Sanger kinase study is by far the largest dataset relating to somatic kinase mutations, and has the advantage of generating a clear picture of the background mutational levels in each gene and in each tumour. This allows the estimation of the likelihood of mutations in a particular

gene being significant as a driver of each particular tumour. The deviation of the ratio of non-synonymous:-synonymous mutations from that expected by chance, was used to indicate the presence of selection in genes on non-synonymous mutations.

Each gene is assigned the selective pressure calculated by Greenman *et al*. (11): Of the 921 base substitutions in the primary screen 763 are estimated to be passenger mutations, with an estimated 158 driver mutations predicted to be distributed within 119 genes. Each gene is also assigned a rank, which reflects the probability of the gene containing at least one driver mutation.

Data normalization and domain assignment

The translated CGP kinase nucleotide sequences were used as reference sequences, and were scanned against Swiss-Prot/TrEMBL (18) using BLASTP (19), to identify the corresponding Swiss-Prot protein entry, which was then used for numbering, annotation and linking to other major primary and secondary databases. Kinase sequences were mapped to the closest Swiss-Prot sequence and the alignments stored in the database—where more than one isoform was identified the longest transcript was adopted as the matched sequence. This protocol was repeated for the reference sequences from the COSMIC database.

Kinase classification schema were extracted from KinBase, and each kinase assigned to its group, family and sub-family. Alternative Gene identifiers were extracted from KinBase and Swiss-Prot and stored in a pseudonym table. Gene names were also mapped to HUGO identifiers (20). Domain boundaries were extracted from Pfam (21), and both PfamA and PfamB domains are displayed. PROSITE patterns (22) are used to identify kinase signature patterns, for example the Serine/Threonine protein kinases active-site signature and Protein kinases ATP-binding region signature. Known phosphorylation sites were downloaded from Phospho.ELM (23) and mapped on to the kinase sequences to help determine whether a mutation changes a residue normally post-translationally modified during its functional life.

Structural mapping of mutations

To map mutational data to protein structures, the sequence for each Pfam domain or non-domain region that contained a missense mutation, was scanned against a database containing non-redundant protein sequences (UniRef90) (18) and the sequences from the PDB using PSI-BLAST (cut-off value of 10^{-4}) (19). The kinase reference sequences and PDB sequences were then aligned.

To identify which mutations mapped onto residues with structural density in the PDB file, PDB sequence to structure alignments from the Structure integration with function, taxonomy and sequence (SIFTS) initiative (24) were utilized. Models of the structures with the substituted residues are currently being generated and will be available shortly.

Automated assessment of pathogenicity of mutations

Prediction of which missense mutations contribute to oncogenesis, are provided by the CanPredict algorithm (12), which incorporates three independent scoring schemes based on SIFT (25), Pfam logR.E (26) and GOSS scores. The SIFT algorithm uses similarity between closely related proteins to identify potentially deleterious changes. SIFT scores <0.05 are predicted to be deleterious. The Pfam-based logR.E-value score predicts whether a change will alter protein function by determining the difference in fit of a wild-type version of the protein to a particular Pfam model and a score >0.5 indicates a deleterious change. Lastly, a GOSS metric uses the gene ontology to measure the similarity of the function of a gene to other known cancer-causing genes.

Protein–protein interaction data

Key to determining the impact of a mutation on the cell, is the impact of the mutation on the proteins interactions and pathways. Towards this goal, protein–protein interaction data, for each kinase are provided by the ROCK database (Zvelebil *et al*., unpublished), an Oracle-based data warehouse that integrates a large number of experimental and derived data in a modular design. ROCK will provide a single online interface for the management, navigation, cross-linking and cross-correlation of data relating to breast cancer research from a wide variety of laboratory and clinical studies. The core of the database contains genomic and protein interaction datasets which are linked to experimental results. Interaction datasets are from MINT (27), MIPS (28), Reactome (29), BioGRID (30), HPRD (31), IntAct (32), BIND (33) and derived from Inparanoid (34) and Homologene (35) along with data curated from literature (Supplementary Table 1).

Database design

MoCKa was implemented using a MySQL database running on a Linux server, with PERL scripts used for all data retrieval and output. Its modular design is compatible with future expansion and connectivity, and currently contains the following subsections: Gene [EntrezGene (35), RefSeq (35), Ensembl (36)], Pseudonyms, Protein (Swiss-Prot/UniProt). Gene Ontology (37), Structure (PDB) (38), Mutational data [CGP, COSMIC (17)], Domain Structure and annotation [CATH (39), Gene3D (40), Pfam (21), SMART (41)], Functional annotation [Phospho.ELM (19)], Cancer data [Sequence, Mutations, Cell line, Cancer sub-type, Selective Pressure (9)] and protein interactions for each kinase (ROCK).

THE MoKCa WEB-INTERFACE

At the highest level MoKCa provides the full list of 518 human protein kinases listed alphabetically by gene name to facilitate browsing, with each entry labelled with the number of mutations found, the cancer driver selection pressure and rank, and an iconic representation of the tumour type(s) in which mutations in that protein kinase have been found. A 'lightbulb' icon indicates those proteins for which expert annotations have been added. In addition to alphabetic sorting on the gene name, the list can be sorted by selective pressure or driver ranking (Figure 1). Sorting on selective pressure is particularly

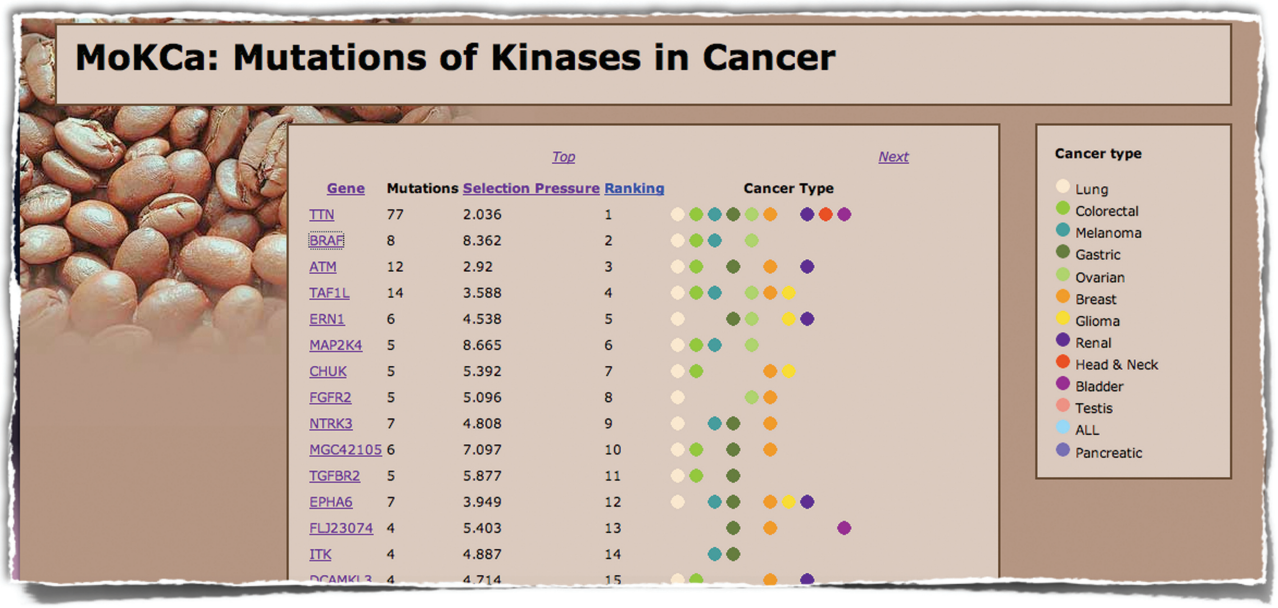

Figure 1. MoKCa kinase gene list. Genes encoding protein kinases are shown listed by ranking of their probability of containing one or more cancer-driving mutation (11). Gene names are additionally annotated with number of mutations found in the Cancer Genome Project analysis (9), the calculated selection pressure on that gene, and indicators showing the cancer types in which the gene was found mutated. The list can also be sorted alphabetically or by selective pressure. Gene names hyperlink to gene-level pages.

informative in presenting those kinases that were found to have the highest involvement in driving cancer in the study of Greenman *et al.* (11), and probably presents the least biased assessment of driver probability currently available. Alternatively, a filtered list can be generated by flexible text-matching against gene name, UniProt accession code, UniProt protein names and synonyms, or GenBank ID. Future developments will allow selection of sub-sets of genes based on attributes such as kinome 'branch' or association with specific tumour type(s). Website navigation is assisted by a 'breadcrumb' system that tracks the user's journey through the database and allows them to return to any stage.

Gene level

Selection of a gene from the full or filtered list transfers the user to a gene-level page, which provides information and links relevant to the gene, its encoded protein and the spectrum of mutations present in the database for that gene. The encoded protein structure is represented as a domain pictogram, with the positions of silent, missense, nonsense, indel and frameshift mutations indicated. Phosphorylation sites and other functional features are also indicated (Figure 2a and b). A 3D-cartoon image of the crystal structure most closely homologous to the encoded protein and to which the greatest number of mutations can be mapped, is shown, with the mapped positions of mutated residues highlighted (see later).

Hyperlinks are provided to entries for the gene in the Cancer Genome Project COSMIC database, SwissProt/UNIPROT and the iHop literature browser (42). Biological function can be accessed via a link to GO and network interactions involving the encoded protein can be explored via a link to ROCK.

The list of individual mutations identified for the gene, annotated by the amino acid change, Pfam domain to which they map, cancer driver prediction from CanPredict ('tick' for probably cancer; 'cross' for probably not) and tumour type(s) in which they occur, provide hyperlinks to the mutation level. The Pfam domain names (for Pfam-A only) hyperlink to the functional definitions for that domain. A 'lightbulb' icon indicates those mutations for which expert annotations have been added.

Mutation level

The mutation-level pages provide information and links relevant to the particular mutation, including details of the tumour sample and type in which it was found, and the breakdown of the CanPredict assessment of the cancer probability of that mutation, where available. SMART and Pfam domain boundary mappings on to the amino acid sequence are also provided. A hyperlinked list is provided of all Protein Databank 3D structures (ranked by sequence identity) that match the amino acid sequence of the affected domain with a significant PSI-BLAST e-value. The list is sorted by sequence identity with each entry annotated with the BLAST e-value for the match, and the protein chain and residue number in the structure file that corresponds to the mutated residue. A 3D-cartoon image is generated of the highest homology structure in which the equivalent to the affected residue was structurally defined, with that residue highlighted. Hyperlinks are provided to launch an interactive session with the Jmol viewing applet (http://www.jmol.org), in which the structure can be examined in 3D, and a script for the fully

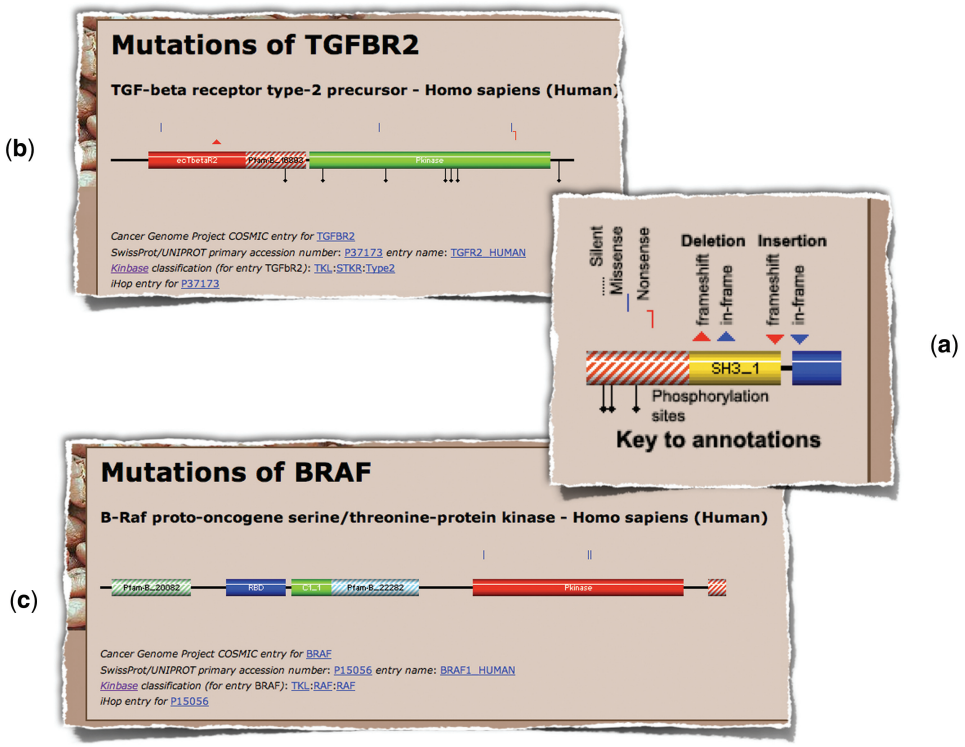

Figure 2. Mapping mutations to domains. The spectrum of mutations identified for each gene is shown on its gene-level page mapped on to a schematic representation of the domain structure of the encoded protein. (**a**) The domain structure of each encoded protein (defined by Pfam definitions) is shown as a pictogram, with the positions of mutations annotated with icons specific to the type of mutation—silent, missense, nonsense (stop), deletion (frameshift or in-frame) or insertion (frameshift or in-frame). Functionally important sequence features, such as documented phosphorylation sites from Phospho.ELM (23) are also shown. Annotation of features such as active sites and ATP-binding motifs (from Prosite) will be added in future versions. (**b**) Example domain pictogram for TGFBR2, showing missense mutations distributed in the extra-cellular and kinase domains, a frameshift (probably truncating) deletion mutation in the extra-cellular domain, and a truncating nonsense mutation in the C-terminal lobe of the kinase domain. Automated assessment by CanPredict identifies a lung-cancer associated H328Y mutation as a cancer driver. (**c**) Domain pictogram for BRAF, identified as a strongly selected mutated gene in melanoma and a range of other cancers, showing the cluster of activating missense mutations in the activation segment.

featured molecular Graphics program PyMol (http://www.pymol.org) can be downloaded.

Registered 'expert' users are able to access text fields in which they can enter their expert assessment of the mechanistic effect of the particular mutation based on the automated predictions and detailed examination of the structural mapping. The mechanism describes the change in biological activity that results from the mutation and includes whether the kinase activity increases or decreases, and whether this results in a tumour suppressor or oncogenic effect on the cell or its pathways. Whereas the evidence refers to the publications in which the proposed mechanism has been described. It can be direct or inferred from a similar mutation in a homologous protein or in an analogous system.

These opinions are visible on the mutation-level pages, but not editable, by other users. Future developments will include hyperlinked bibliographic references in support of the expert opinion.

Protein functional interaction networks

Interaction maps are displayed using ROCKscape with nodes representing proteins and edges representing interaction or transcriptional regulatory relationships between components. Depending on the respective display mode selected, edges are colour-coded to identify the type of interaction, and nodes are coloured with selective pressure. Interaction maps can also be progressively expanded outwards to display the next interaction neighbourhood or subsequent levels beyond, thus, making it possible to 'grow' networks or extend signalling pathway modules in a desired direction.

EXAMPLE OF USE

One of the first discoveries of the Cancer Genome Project was that mutations in BRAF were found in 70% of melanomas, and to a lesser extent in colorectal, lung and ovarian cancers. By browsing the MOKCA interface, and choosing 'selective pressure', it is clear the BRAF is near the top of the predicted driver list (Figure 1), and it can be easily identified in which tumour types the mutations arise. By following the link to the protein page it is clearly visible that the mutations are all missense (Figures 2c and 3a), are localized to the kinase domain, and most are predicted to be pathogenic by CanPredict. The illustration of the protein structure clearly shows that the mutations are all tightly clustered in the 3D structure,

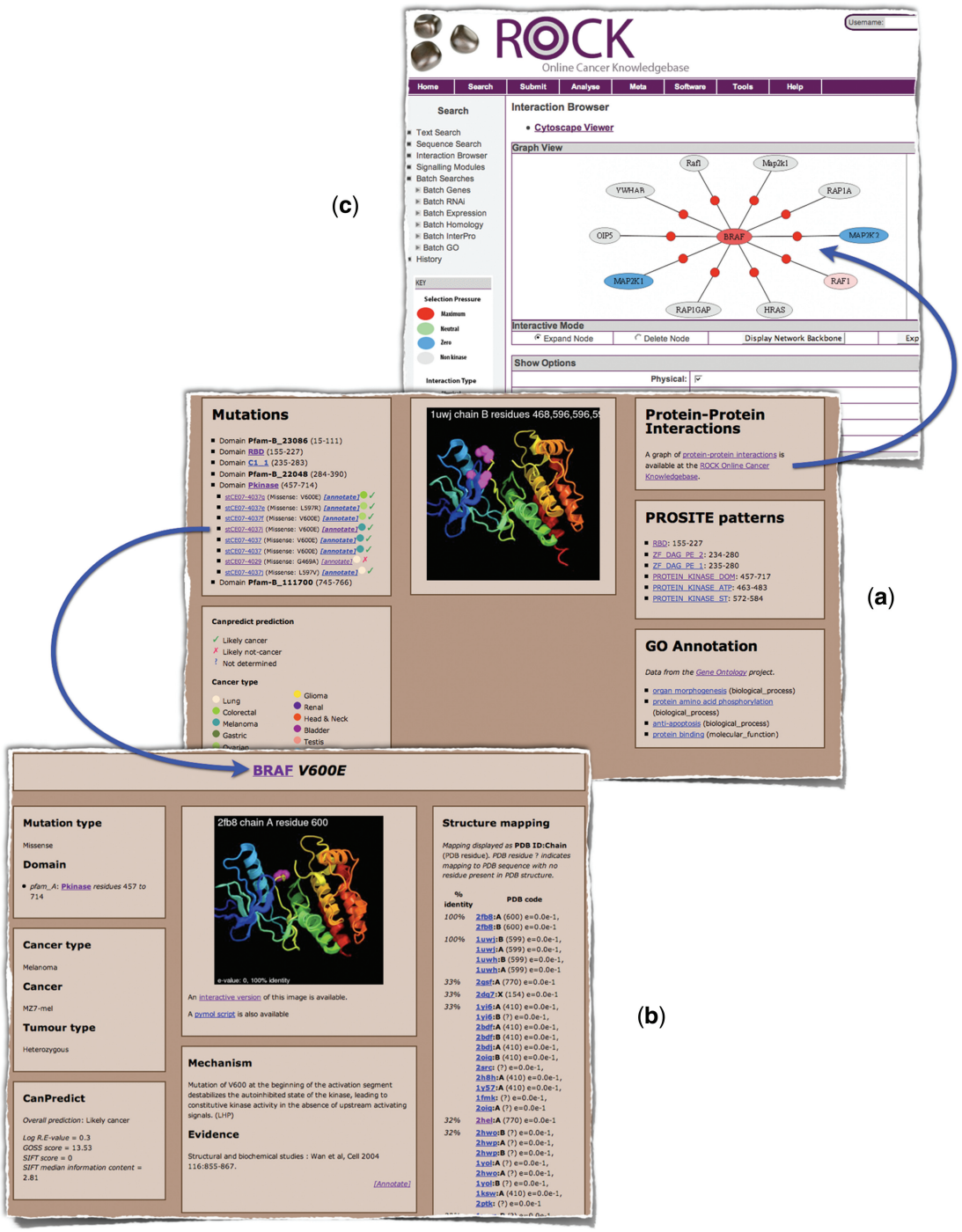

Figure 3. Mutations, mechanisms and pathways. (**a**) Gene-level pages (illustrated for BRAF) provide links to data for the gene in a range of primary and other databases including COSMIC, UNIPROT and KinBase. Additional links are provided to GO functional annotations, Pfam definitions for domains of the encoded protein, and Prosite motifs matched in the encoded protein sequence. A hyperlink to the ROCK Online Cancer Knowledge base (Zvelebil *et al.*, unpublished), provides view of the protein–protein interactions [see (c)]. A hyperlinked list of mutations, annotated with tumour sample registry numbers, and indicators for tumour type and CanPredict assessment of driver probability, points out to individual mutation-level pages [see (b)]. (**b**) Mutation-level pages give specific domain assignments and CanPredict assessments for the individual mutation, as well as hyperlinks to a homology-ranked list of protein structures homologous to the affected domain. A 3D-graphic overview of the domain is provided with the mutated residue highlighted, and a Jmol session allowing interaction with the protein can be launched. 'Expert' users are able to enter a description of the mechanistic implications of the particular mutation, if any, and evidence for that opinion. Future developments will include bibliographic hyperlinks to relevant publications. (**c**) ROCK provides a graphical view (with optional Cytoscape viewer) of the documented protein–protein interactions for the affected gene (highlighted red). For other kinases appearing in the interaction network, the selective pressure for cancer-associated mutations is used to colour the node. In the BRAF example shown, the two ERK kinases (MAP2K1/2) downstream of BRAF display no selective pressure, whereas RAF1 (aka C-RAF) which co-operates in parallel with BRAF, displays a selective pressure for cancer driver mutations. The displayed network can be interactively expanded or contracted, and interrogated, as required. A full description of ROCK will be presented elsewhere.

in and around the activation segment and phosphate-binding loop.

Moving to the mutation pages, an annotated mechanism for each mutation is described (Figure 3b). For instance, Val 600, which is mutated to glutamate in several different melanomas as well as in ovarian and colorectal tumours, lies at the beginning of the activation segment. The V600E mutation destabilizes the auto-inhibited state of the kinase, leading to constitutive kinase activity in the absence of upstream activating signals. This provides the tumours with self-sufficiency in growth signals, one of the key 'hallmark' traits of cancer (4) via its activation of the downstream MAP kinase signalling pathway.

Following a hyperlink to the protein–protein interaction page provided by ROCK (Figure 3c), the functional interaction of BRAF with its downstream kinase phosphorylation targets ERK1/2 (MAP2K1/2) are evident. Unlike the upstream BRAF, neither of these were found to be under selective pressure for mutation in the tumours analysed. It can also be seen that BRAF interacts with RAF1, a known kinase proto-oncogene that also phosphorylates ERK1/2, and like BRAF is also under mutational selective pressure.

DISCUSSION

MoKCa provides data and assessments from multiple sources and algorithms for each potential cancer-associated mutation, and the interactive display of the data facilitates authoritative annotation by cancer biologists and structural biologists. These 'experts' will bring detailed insights into the biochemistry and biology of individual proteins and systems, that are virtually impossible to encapsulate in an algorithm, but are key to determining if and how a particular mutation will alter the biological function of protein. Thus the MoKCa database combines automated and 'expert' annotation of individual mutations, and is firmly directed towards the specific needs of the cancer research community.

FUTURE DEVELOPMENTS

As the various cancer genome and re-sequencing projects continue to generate mutational data, these will be incorporated into the database. We also propose to add cancer-related germline mutational data from OMIM, and family-specific mutational databases such as SH2base (15) and KinMutBase (16), which document inherited genetic disease mutations.

The MoKCa database has been designed to combine functional and protein-interaction data, with a user-friendly interface to assess which kinases are most causal of cancer, and which mutations are cancer drivers. It will be important to be able to mine this information to predict which mutated kinases will be effective drug-targets.

The database will also be extended to include pathway information, for both known and in-house pathways. For fully documented pathways we will try to ascertain at the molecular level, what affect a mutation will have on pathway function and how it affects the 'sign' (activatory or inhibitory) of the pathway interactions made by the affected protein.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Mike Stratton and Erin Pleasance for providing data, Christine Orengo for helpful suggestions and Roman Laskowski for help with (and code for) the domain pictograms.

FUNDING

Funding for open access charge: The Institute of Cancer Research.

Conflict of interest statement. None declared.

REFERENCES

1. Daley,G.Q., Van Etten,R.A. and Baltimore,D. (1990) Induction of chronic myelogenous leukemia in mice by the P210bcr/abl gene of the Philadelphia chromosome. *Science*, **247**, 824–830.
2. Burnworth,B, Arendt,S., Muffler,S., Steinkraus,V., Bröcker,E.B., Birek,C., Hartschuh,W., Jauch,A. and Boukamp,P. (2007) The multi-step process of human skin carcinogenesis: a role for p53, cyclin D1, hTERT, p16, and TSP-1. *Eur. J. Cell. Biol.*, **86**, 763–780.
3. Koorstra,J.B., Hustinx,S.R., Offerhaus,G.J. and Maitra,A. (2008) Pancreatic Carcinogenesis. *Pancreatology*, **8**, 110–125.
4. Hanahan,D and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.
5. Sawyers,C. (2004) Targeted cancer therapy. *Nature*, **432**, 294–297.
6. Manning,G., Whyte,D.B., Martinez,R., Hunter,T. and Sudarsanam,S. (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.
7. Braconi Quintaje,S and Orchard,S. (2008) The annotation of both human and mouse kinomes in UniProtKB/Swiss-Prot: one small step in manual annotation, one giant leap for full comprehension of genomes. *Mol. Cell Proteomics*, **7**, 1409–1419.
8. Sjöblom,T., Jones,S., Wood,L.D., Parsons,D.W., Lin,J., Barber,T.D., Mandelker,D., Leary,R.J., Ptak,J., Silliman,N. *et al.* (2006) The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**, 268–274.
9. Greenman,C., Stephens,P., Smith,R., Dalgliesh,G.L., Hunter,C., Bignell,G., Davies,H., Teague,J., Butler,A., Stevens,C. *et al.* (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446**, 153–158.
10. Wood,L.D., Parsons,D.W., Jones,S., Lin,J., Sjöblom,T., Leary,R.J., Shen,D., Boca,S.M., Barber,T., Ptak,J. *et al.* (2007) The genomic landscapes of human breast and colorectal cancers. *Science*, **318**, 1108–1113.
11. Greenman,C., Wooster,R., Futreal,P.A., Stratton,M.R. and Easton,D.F. (2006) Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics*, **173**, 2187–2198.
12. Kaminker,J.S., Zhang,Y., Watanabe,C. and Zhang,Z. (2007) CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Res.*, **35**, W595–W598.
13. Torkamani,A. and Schork,N.J. (2008) Prediction of cancer driver mutations in protein kinases. *Cancer Res.*, **68**, 1675–1682.
14. Kaminker,J.S., Zhang,Y., Waugh,A., Haverty,P.M., Peters,B., Sebisanovic,D., Stinson,J., Forrest,W.F., Bazan,J.F., Seshagiri,S. *et al.* (2007) Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer Res.*, **67**, 465–473.
15. Lappalainen,I., Thusberg,J., Shen,B. and Vihinen,M. (2008) Genome wide analysis of pathogenic SH2 domain mutations. *Proteins*, **72**, 779–792.

16. Ortutay,C., Väliaho,J., Stenberg,K. and Vihinen,M. (2005) KinMutBase: a registry of disease-causing mutations in protein kinase domains. *Hum. Mutat.*, **25**, 435–442.
17. Forbes,S.A., Bhamra,G., Bamford,S., Dawson,E., Kok,C., Clements,J., Menzies,A., Teague,J.W., Futreal,P.A. and Stratton,M.R. (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC).*Curr. Protoc. Hum. Genet.* Chapter 10:Unit 10.11.
18. UniProt Consortium. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.
19. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
20. Bruford,E.A., Lush,M.J., Wright,M.W., Sneddon,T.P., Povey,S. and Birney,E. (2008) The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res.*, **36**, D445–D448.
21. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
22. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., Cuche,B.A., de Castro,E., Lachaize,C., Langendijk-Genevaux,P.S. and Sigrist,C.J. (2008) The 20 years of PROSITE. *Nucleic Acids Res.*, **36**, D245–D249.
23. Diella,F., Gould,C.M., Chica,C., Via,A. and Gibson,T.J. (2008) Phospho.ELM: a database of phosphorylation sites–update 2008. *Nucleic Acids Res.*, **36**, D240–D244.
24. Velankar,S., McNeil,P., Mittard-Runte,V., Suarez,A., Barrell,D., Apweiler,R. and Henrick,K. (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33**, D262–D265.
25. Ng,P.C. and Henikoff,S. (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Res.*, **12**, 436–446.
26. Clifford,R.J., Edmonson,M.N., Nguyen,C. and Buetow,K.H. (2004) Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics*, **20**, 1006–1014.
27. Chatr-aryamontri,A., Ceol,A., Palazzi,L.M., Nardelli,G., Schneider,M.V., Castagnoli,L. and Cesareni,G. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **35**, D572–D574.
28. Pagel,P., Kovac,S., Oesterheld,M., Brauner,B., Dunger-Kaltenbach,I., Frishman,G., Montrone,C., Mark,P., Stümpflen,V., Mewes,H.W., Ruepp,A. and Frishman,D. (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics*, **21**, 832–834.
29. Vastrik,I., D'Eustachio,P., Schmidt,E., Joshi-Tope,G., Gopinath,G., Croft,D., de Bono,B., Gillespie,M., Jassal,B., Lewis,S. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.
30. Stark,C., Breitkreutz,B.J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: A General Repository for Interaction Datasets. *Nucleic Acids Res.*, **34**, D535–D539.
31. Mishra,G.R., Suresh,M., Kumaran,K., Kannabiran,N., Suresh,S., Bala,P., Shivakumar,K., Anuradha,N., Reddy,R., Raghavan,T.M. *et al.* (2006) Human Protein Reference Database - 2006 Update. *Nucleic Acids Res.*, **34**, D411–D414.
32. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. *et al.* (2007) IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
33. Alfarano,C., Andrade,C.E., Anthony,K., Bahroos,N., Bajec,M., Bantoft,K., Betel,D., Bobechko,B., Boutilier,K., Burgess,E. *et al.* (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.*, **33**, D418–D424.
34. O'Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.
35. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
36. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.
37. Gene Ontology Consortium (2008) The Gene Ontology project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.
38. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
39. Pearl,F., Todd,A., Sillitoe,I., Dibley,M., Redfern,O., Lewis,T., Bennett,C., Marsden,R., Grant,A., Lee,D. *et al.* (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.*, **33**, D247–D251.
40. Yeats,C., Lees,J., Reid,A., Kellam,P., Martin,N., Liu,X. and Orengo,C. (2008) Gene3D: comprehensive structural and functional annotation of genomes. *Nucleic Acids Res.*, **36**, D414–D418.
41. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.
42. Fernández,J.M., Hoffmann,R. and Valencia,A. (2007) iHOP web services. *Nucleic Acids Res.*, **35**, W21–W26.